Random Distanceの分布を用いるクラスター分析 (多次元統計解析の数理的研究)

AUTHOR(S):

脇本, 和昌; 山本, 英二; 垂水, 共之

数理解析研究所講究録
345巻 1979年 1-6

Random Distanceの分布を用いるクラスター分析

岡山大 養　　脇本 和昌
岡山理大　　山本 英二
岡山大 養　　垂水 共之

[1] Random Distanceの分布

互いに独立で、(0,1)の一様分布に従うk個の要素で構成されたk次元random vector $(X_1, X_2, \cdots\cdots, X_k)$を考える。2つのrandom vector $(X_1, X_2, \cdots\cdots, X_k)$, $(X_1', X_2', \cdots\cdots, X_k')$の距離 $D_k=\sqrt{\sum_1^k (X_i-X_i')^2}$ をrandom distanceと呼ぶことにする。random distanceの2乗D_k^2の$k=1,2,3$のとき分布関数$F_k(d^2)$は以下の通り、

$$F_1(d^2)=\begin{cases} 0 & d^2\leqq 0 \\ -d^2+2d & 0<d^2\leqq 1 \\ 1 & 1\leqq d^2 \end{cases}$$

$$F_2(d^2)=\begin{cases} 0 & d^2\leqq 0 \\ \pi d^2-\frac{8}{3}d^3+\frac{1}{2}d^4 & 0<d^2\leqq 1 \end{cases}$$

$$
\begin{cases}
\frac{1}{3}+(\pi-2)d^2+4(d^2-1)^{\frac{1}{2}}+\frac{8}{3}(d^2-1)^{\frac{3}{2}}-\frac{d^2}{2}-4d^2\sec^{-1}d & 1<d^2\leqq 2 \\
1 & 2<d^2
\end{cases}
$$

$$
F_3(d^2)=\begin{cases}
0 & d^2\leqq 0 \\
\frac{4}{3}\pi d^3-\frac{3}{2}\pi d^4+\frac{8}{5}d^5-\frac{1}{6}d^6 & 0<d^2\leqq 1 \\
(\frac{5}{2}\pi+\frac{43}{30})-6(d^2-1)^{\frac{1}{2}}+(3\pi+\frac{7}{2})(d^2-1) & \\
\quad -\frac{8}{3}\pi d^3-10(d^2-1)^{\frac{3}{2}}+\frac{5}{2}(d^2-1)^2 & \\
\quad -\frac{16}{5}(d^2-1)^{\frac{5}{2}}+\frac{1}{3}(d^2-1)^3+6d^4\sec^{-1}d & 1<d^2\leqq 2 \\
(\frac{23}{2}\pi-\frac{343}{30})+14(d^2-2)^{\frac{1}{2}}+(9\pi-\frac{21}{2})(d^2-2)+10(d^2-2)^{\frac{3}{2}} & \\
\quad +(\frac{3\pi-5}{2})(d^2-2)^2+\frac{8}{5}(d^2-2)^{\frac{5}{2}}-\frac{1}{6}(d^2-2)^3 & \\
\quad -(6d^4+12d^2-2)\sec^{-1}\sqrt{d^2-1}+8d^3\sec^{-1}(d^2-1)-\frac{8}{3}\pi d^3 & 2<d^2\leqq 3 \\
1 & 3<d^2
\end{cases}
$$

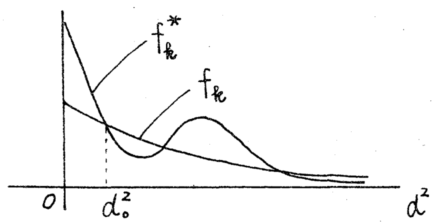
kが大きいときの D_k^2 の漸近分布は、正規分布であるが、原点の回りのr次のモーメントが $\frac{1}{(2r+1)(r+1)}$ であることを利用して、Gram-Charlier-Edgeworth 近似によって、近似の精度を上げることが出来る。

〔2〕 クラスター分析のアルゴリズム

1. k次元単位立方体上のN個(N≫1)のデータが作る $_NC_2$

個の線分から復元抽出により大きさnのランダムサンプルを抽出し、それにもとずき距離の経験分布 F_k^* を計算する。

2. 〔1〕で求めた分布F_kとの適合度検定を行い、有意水準α%で棄却されたとき、クラスターがあると判断し、さらに下図の様にd_0を定め、以下、クラスター(d_0)を求めていく。



3. 単位立方体の中に一辺 $d_0' = \frac{1}{\sqrt{k}} \cdot \frac{1}{\sqrt[k]{M}} \cdot d_0$ ($M = 1 + \frac{\log_{10} N}{\log_{10} 2}$：スタージェスの式)の立方体を系統的に$m$個とり($S_1, \cdots\cdots, S_m$)、$S_i$に含まれる全ての標本点の個数を$l_i$とし、$l_i$個の点の密度と${}_{l_i}C_2$個の距離の経験分布を計算する。

4. 各S_iにおいて、密度と経験分布を用いてクラスターがあるかないか判断し、ある場合は、重心を seed point とする。

5 クラスターを有するS_iの結合を seed point 間距離と $d_0'' = \frac{1}{\sqrt[k]{M}} d_0$ の比較で行う。

〔3〕 この手法の特徴

1. クラスターの決定を自動的に行う。2点間距離の経験分布を使って、クラスターを定義し、それを求めていくことになる。
2. クラスターの判断を密度だけでなく、2点間距離の経験分布を用いて行っていること。これにより、S_iにおいて、高密度でなくても、点が偏在しているときは、クラスターと判断可能となる。
3. 大標本への適用可能

　　サンプリングと、立方体の分割により、計算量を大巾に減らすことができる。

〔4〕 適用例

図1に示すような2つのクラスターと1つの離れた点からなる2変量のデータを考える。このデータに対してここで提案した方法を用いると、左上5つのseed pointと右下4つのseed pointができて、そのseed pointの結合により2つのクラスターと1つの離れた点とに分けることができた。なお1辺d_0'の正方形の内部に1点しかはいらない場合は、1番近いseed pointまでの距離がd_0''内であれば、その点をseed pointに属させてクラスターを作った。

図 1

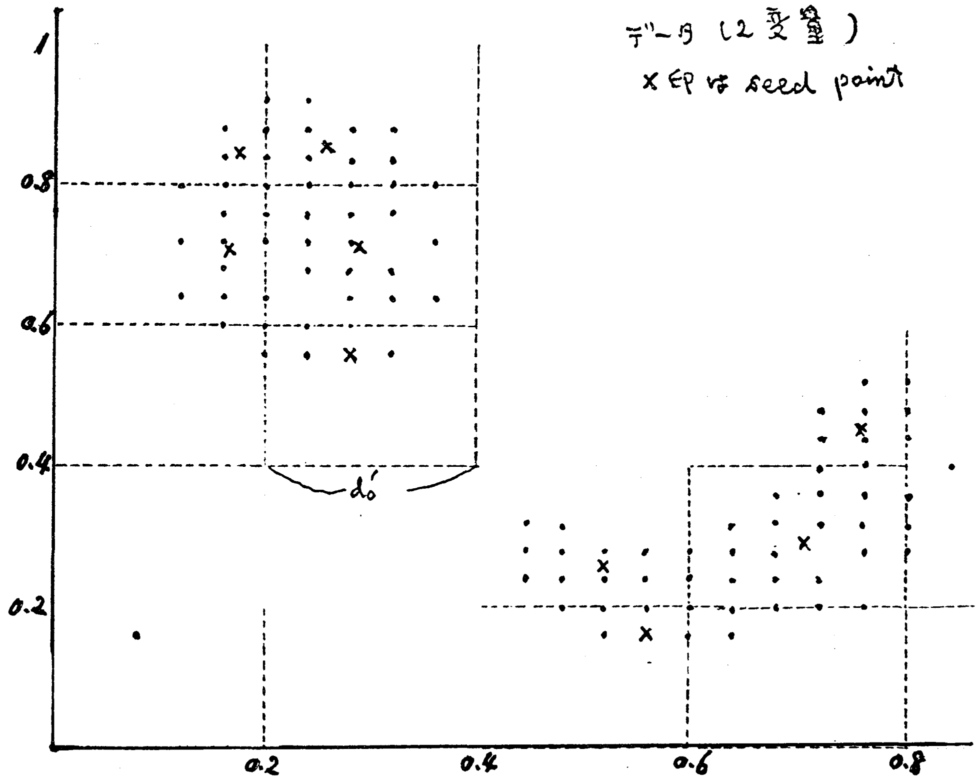
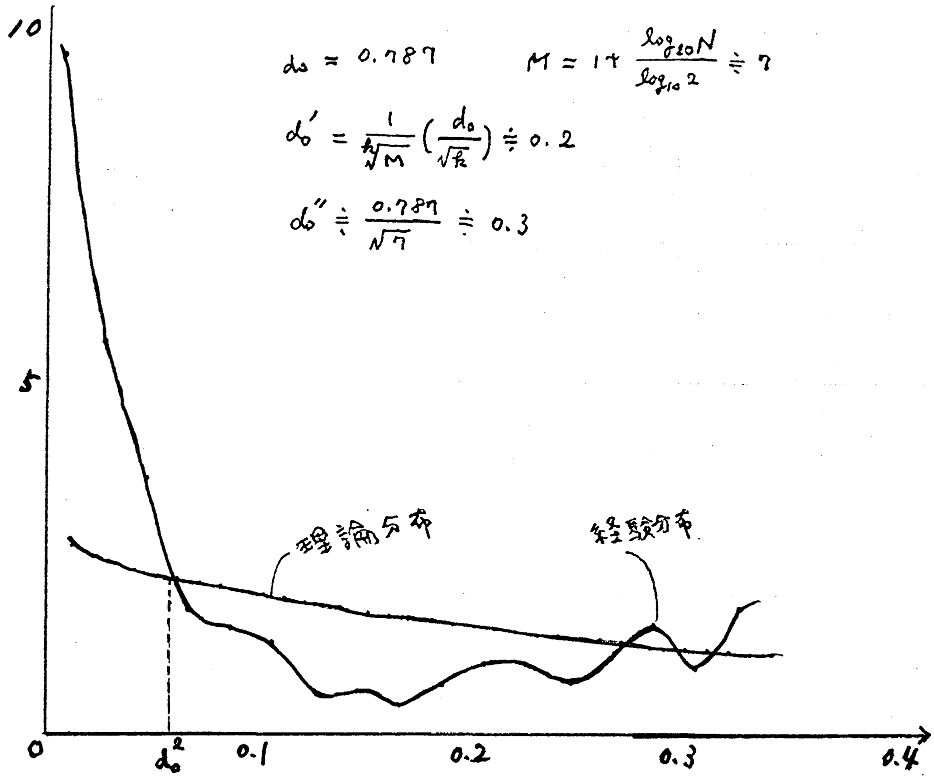

データ (2変量)
×印は seed point
1
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
d_0'

図 2


10
5
0
d_0^2
0.1
0.2
0.3
0.4
$d_0 = 0.787$
$M = 1 + \frac{\log_{10} N}{\log_{10} 2} \doteqdot 7$
$d_0' = \frac{1}{\sqrt[k]{M}}\left(\frac{d_0}{\sqrt{k}}\right) \doteqdot 0.2$
$d_0'' \doteqdot \frac{0.787}{\sqrt{7}} \doteqdot 0.3$
理論分布
経験分布

今後の研究課題としては、他のクラスター分析の手法との比較検討をおこなうこと、また複雑な例題についてその効果を例示することを考えている。

参考文献

[1] F. Gruenberger & A. M. Mark (1951) "The χ^2-test of random digits" Math. Tables other Aids Comp. 5. pp. 109 − 110.

[2] A. A. TÖRN (1977) "Cluster analysis using seed points and density-determined hypersheres as an aid to global optimization" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBENETICS, Vol. SMC-7, NO-8, AUGUST pp. 610 − 616.